Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu

Abstract—We introduce a new dataset, Human3.6M, of 3.6 Million accurate 3D Human poses, acquired by recording the performance of 5 female and 6 male subjects, under 4 different viewpoints, for training realistic human sensing systems and for evaluating the next generation of human pose estimation models and algorithms. Besides increasing the size of the datasets in the current state-of-the-art by several orders of magnitude, we also aim to complement such datasets with a diverse set of motions and poses encountered as part of typical human activities (taking photos, talking on the phone, posing, greeting, eating, etc.), with additional synchronized image, human motion capture, and time of flight (depth) data, and with accurate 3D body scans of all the subject actors involved. We also provide controlled mixed reality evaluation scenarios where 3D human models are animated using motion capture and inserted using correct 3D geometry, in complex real environments, viewed with moving cameras, and under occlusion. Finally, we provide a set of large-scale statistical models and detailed evaluation baselines for the dataset illustrating its diversity and the scope for improvement by future work in the research community. Our experiments show that our best large-scale model can leverage our full training set to obtain a 20% improvement in performance compared to a training set of the scale of the largest existing public dataset for this problem. Yet the potential for improvement by leveraging higher capacity, more complex models with our large dataset, is substantially vaster and should stimulate future research. The dataset together with code for the associated large-scale learning models, features, visualization tools, as well as the evaluation server, is available online at http://vision.imar.ro/human3.6m.

Index Terms—3D human pose estimation, human motion capture data, articulated body modeling, optimization, large-scale learning, structured prediction, Fourier kernel approximations

1 Introduction

Accurately reconstructing the 3D human poses of people from real images, in a variety of indoor and outdoor scenarios, has a broad spectrum of applications in entertainment, environmental awareness, or human-computer interaction [1]–[3]. Over the past 15 years the field has made significant progress fueled by new optimization and modeling methodology, discriminative methods, feature design and standardized datasets for model training. It is now widely agreed that any successful human sensing system, be it generative, discriminative or combined, would need a significant training component, together with strong constraints from image measurements, in order to be successful, particularly under monocular viewing and (self-) occlusion. Such situations are not infrequent but rather commonplace in the analysis of images acquired in real world situations. Yet these images cannot be handled well with the human models and training tools currently available in computer vision. Part of the problem is that humans are highly flexible, move in complex ways against natural backgrounds, and their clothing and muscles deform. Other confounding factors like occlusion may also require comprehensive scene modeling, beyond just the humans in the scene. Such image understanding scenarios stretch the ability of the pose sensing system to exploit prior knowledge and structural correlations, by using the incomplete visible information in order to constrain estimates of unobserved body parts. One of the key challenges for trainable systems is insufficient data coverage. Existing state of the art datasets like HumanEva [4], contain about 40,000 different poses and the class of motions covered is somewhat small, reflecting its design purpose geared primarily towards algorithm evaluation. In contrast, while we want to continue to be able to offer difficult benchmarks, we also wish to collect datasets that can be used to build operational systems for realistic environments. People in the real world move less regularly than assumed in many existing datasets. Consider the case of a pedestrian, for instance. It is not that frequent, particularly in busy urban environments, to encounter 'perfect' walkers. Driven by their daily tasks, people carry bags,

- *C. Ionescu is with the Institute of Mathematics of the Romanian Academy (IMAR), Bucharest RO010702, Romania, and also with the Faculty of Mathematics and Natural Sciences, University of Bonn, Bonn D53115, Germany. E-mail: catalin.ionescu@ins.uni-bonn.de.*
- *D. Papava and V. Olaru are with the Institute of Mathematics of the Romanian Academy (IMAR), Bucharest RO010702, Romania. E-mail: {dragos.papava; vlad.olaru}@imar.ro.*
- *C. Sminchisescu is with the Department of Mathematics, Faculty of Engineering, Lund University, Lund SE22100, Sweden, and also with IMAR. E-mail: cristian.sminchisescu@math.lth.se.*
- *C. Ionescu and D. Papava contributed equally. Corresponding authors: V. Olaru, C. Sminchisescu.*

Manuscript received 16 Mar. 2013; revised 1 Sep. 2013; accepted 12 Nov. 2013. Date of publication 11 Dec. 2013; date of current version 13 June 2014. Recommended for acceptance by M. Pantic.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier 10.1109/TPAMI.2013.248

walk with hands in their pockets and gesticulate when talking to other people or on the phone. Since the human kinematic space is too large to be sampled regularly and densely, we chose to collect data by focusing on a set of poses which are likely to be of interest because they are common in urban and office scenes. The poses are derived from 15 chosen scenarios for which our actors were given general instructions, but were also left ample freedom to improvise. This choice helps us cover more densely some of the common pose variations and at the same time control the difference between training and testing data (or covariate shift [5]) without placing unrealistic restrictions on their similarity. However that variability *within* daily tasks like "talking on the phone" or "Eating" is subtle as functionally, similar programs are being performed, irrespective of the exact execution. In contrast, the distributions of any two such *different* scenarios are likely to contain wider separated poses, although the manifolds from which this data is sampled may intersect.

In this paper we present a large dataset collected using accurate marker-based motion capture systems and actors dressed with moderately realistic clothing, viewed against indoor backgrounds. Other recent experimental systems have explored the possibility of unconstrained capture based on non-invasive sensors or attached body cameras [6], [7], and could represent attractive alternatives, as they develop, in the long run. As technology matures, progress on all fronts is welcome, particularly as the data we provide is complementary in its choice of poses, with respect to existing datasets [4], [7], [8]. Even by means of a combined community effort, we are not likely to be able to densely sample or easily handle the 30+ dimensional space of all human poses. However, an emphasis on typical scenarios and larger datasets, in line with current efforts in visual recognition, may still offer a degree of prior knowledge bootstrapping that can significantly improve the performance of existing human sensing systems. Specifically, by design, we aim to cover the following aspects:

Large Set of Human Poses, Diverse Motion and Activity Scenarios: We collected over 3.6 million different human poses, viewed from 4 different angles, using an accurate human motion capture system. The motions were executed by 11 professional actors, and cover a diverse set of everyday scenarios including conversations, eating, greeting, talking on the phone, posing, sitting, smoking, taking photos, waiting, walking in various non-typical scenarios (with a hand in the pocket, talking on the phone, walking a dog, or buying an item). Fig. 1 shows an example of the type of poses we focus on as well as images of actors reproducing such poses and the corresponding 3D reconstructions.

Synchronized Modalities, 2D and 3D data, Subject Body Scans: We collect and fully synchronize both the 2D and the 3D data, in particular images from 4 high-speed progressive scan, high-resolution video cameras, a time of flight (TOF) depth sensor, as well as human motion capture data acquired by 10 high-speed cameras. We also provide 3D full body models of all subjects in the dataset, acquired with an accurate 3D laser scanner.

Evaluation Benchmarks, Complex Backgrounds, Occlusion: The dataset provides not only training, validation and testing sources for the data collected in the laboratory, but also a variety of mixed-reality settings where realistic graphical characters have been inserted in video environments collected using real, moving digital cameras, and animated using our motion capture data. The insertions and occlusions are geometrically correct, based on estimates of the camera motion and its internal parameters, reconstructions of the 3D environment and ground plane estimates.

Online Large-Scale Models, Features, Visualization and Evaluation Tools: We provide online models for feature extraction as well as pose estimation, including linear and kernel regressors and structured predictors based on kernel dependency estimation. All these models are complemented with linear Fourier approximations, in order to allow the training of non-linear kernel models at large scale. The design of such models is currently non-trivial and the task of processing millions of images and 3D poses, or training using such large repositories, remains daunting for most existing human pose estimation methodologies. We also supply methods for background subtraction and for extracting the bounding boxes of people, as well as a variety of *precomputed features* (pyramids of SIFT grids) over these, in order to allow rapid prototyping, experimentation, and parallel work streams in both computer vision and machine learning. Software for the visualization of skeleton representations based on 3D joint positions as well as 3D joint angle formats is provided, too.

1.1 Related Work

Over the past decade, inferring the 3D human pose from images or video has received significant attention in the research community. While a comprehensive survey would be impossible, we refer the reader to recently edited volumes by Moeslund *et al*. [1] and Rosenhahn *et al*. [2] as well as [3], [9] for a comprehensive overview. Initially, work in 3D human sensing focused on 3D body modeling and relied on non-linear optimization techniques. More recently, the interest shifted somewhat towards systems where components are trained based on datasets of human motion capture. Within the realm of 3D pose inference, some methods focus on automatic discriminative prediction [10]–[14], whereas others aim at model-image alignment [15]–[23] or accurate modeling of 3D shape [24], [25] or clothing [26]. This process is ongoing and was made possible by the availability of 3D human motion capture [4], [8], as well as human body scan datasets like the commercially available CAESAR, or smaller academic repositories like SCAPE [27] and INRIA4D [28]. Training models for 3D human sensing is not straightforward, however. The CMU dataset [8] contains a diverse collection of human poses, yet these are not synchronized with the image data, making end to end training and performance evaluation difficult. Due to difficulties in obtaining accurate 3D pose information with synchronized image data, evaluations were initially qualitative. Quantitative evaluations were pursued later using graphic renderings of synthetic models [12], [29], [30]. The release of high-quality synchronized data in the HumaEva benchmark [4] has represented a significant step forward, but its size and pose diversity remain somewhat small. Commercial datasets of human body scans like CAESAR are comprehensive and

Fig. 1. Real image showing multiple people in different poses (left), and a matching sample of our actors in similar poses (middle) together with their reconstructed 3D poses from the dataset, displayed using a synthetic 3D model (right). The desire to cover the diversity of 3D poses present in such real-world environments has been one of our motivations for the creation of **Human3.6M**.

offer statistically significant body shape variations of an entire population, but provide no motion or corresponding image data for the subjects involved. An ambitious effort to obtain 3D pose information by manual annotation was pursued in [31], although the 2D and 3D labelings are only qualitative, and the size of the gathered data is small: 1000 people in 300 images. Approaches to 2D human body localization have also been pursued [31]–[34]. For 2D pose estimation, ground truth data can be obtained by simply labeling human body parts in the image. Existing datasets include stickmen annotations [33], [35] and extensions of poselets with 2D annotations [36].

The advances on various fronts, 2D and 3D, both in terms of methodology and data availability, have motivated the recent interest towards realistic 3D human motion capture in natural environments [6], [7], [37], [38]. Very encouraging results have been obtained, but there are still challenges that need to be solved before the technology will enable the acquisition of millions of human poses. Recent interest in 3D motion capture technologies has been spurred by the public availability of time-of-flight, infrared or structured light sensors [39], [40]. The most well-known of these, the Kinect system, represents a vivid illustration of a successful real-time pose estimation solution deployed in a commercial setting. Its performance is in part due to a large scale training set of roughly 1 million pose samples, which remains proprietary, and in part due to the availability of depth information that simplifies the segmentation of the person from its surroundings, and limits 3D inference ambiguities for limbs. By its size and complexity **Human3.6M** is meant to provide the research community with data necessary to achieve similar performance in the arguably more difficult case of only working with intensity images, or alternatively–through our time-of-flight data–, in similar setups as Kinect, by means of open access and larger and more diverse datasets.

2 Dataset Collection and Design

In this section we describe the capture space and the recording conditions, as well as our dataset composition and its design considerations.

2.1 Experimental Setting

Our laboratory setup, represented in Fig. 2(c), lets us capture data from 15 sensors (4 digital video cameras, 1 time-of-flight sensor, 10 motion cameras), using hardware and software synchronization (see Fig. 2(b) for details). The designated laboratory area is about 6m x 5m, and within it we obtain a region of approximately 4m x 3m of effective capture space, where subjects were fully visible in all video cameras. Digital video (DV) cameras (4 units) are placed in the corners of the effective capture space. A time-of-flight sensor (TOF) is also placed on top of one of the digital cameras. A set of 10 motion capture (MoCap) cameras are rigged on the walls to maximize the effective experimentation volume, 4 on each left and right edge and 2 roughly mid-way on the horizontal edges. A 3D laser body scanner from Human Solutions (Vitus LC3) was used to obtain accurate 3D volumetric models for each of the actors participating in the experiments.

The 3D motion capture system relies on small reflective markers attached to the subject's body and tracks them over time. Tracking maintains the label identity and propagates it through time from an initial pose which is labeled either manually or automatically. A fitting process uses the position and identity of each of the body labels, as well as proprietary human motion models, to infer accurate pose parameters.

2.2 Dataset Structure

In this section we describe the choice of human motions captured in the dataset, the output data types provided as well as the image processing and input annotations that we pre-compute. An example of the data modalities provided in our dataset is shown in Fig. 3.

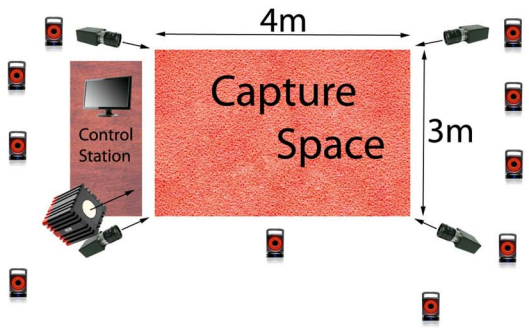
Actors and Human Pose Set: The motions in the dataset were performed by 11 professional actors, 5 female and 6 male, chosen to span a body mass index (BMI) ranging from 17 to 29. We have reserved 7 subjects, 3 female and 4 male, for training and validation, and 4 subjects (2 female and 2 male) for testing. This choice provides a moderate amount of body shape variability as well as different ranges of mobility. Volumetric information in the form of 3D body scans was gathered for each actor to complement the joint position information alone. This data can be used also to evaluate human body shape estimation algorithms [24]. The meshes are released as part of the dataset (see Fig. 5 for examples). The subjects wore their own regular clothing, as opposed to special motion capture outfits, to maintain as much realism as possible. The actors were given detailed tasks and were shown visual examples (images of people) in order to help them plan a stable set of poses for the creation of training, validation and test sets. However, when executing these tasks, the actors were given quite a bit

Type of action	Scenarios	Train	Validation	Test
Upper body movement	Directions	83,856	50,808	114,080
	Discussion	154,392	68,640	140,764
Full body upright variations	Greeting	69,984	33,096	84,980
	Posing	70,948	25,800	85,912
	Purchases	49,096	33,268	48,496
	Taking Photo	67,152	38,216	89,608
	Waiting	98,232	54,928	123,432
Walking variations	Walking	114,468	47,540	93,320
	Walking Dog	77,068	30,648	59,032
	Walking Pair	76,620	36,876	52,724
Variations while seated on a chair	Eating	109,360	39,372	97,192
	Phone Talk	132,612	39,308	92,036
	Sitting	110,228	46,520	89,616
	Smoking	138,028	50,776	85,520
Sitting on the floor	Sitting Down	112,172	50,384	105,396
Various Movements	Miscellaneous	-	-	105,576
Total		1,464,216	646,180	1,467,684

(a) The number of 3D human poses in Human3.6M in training, validation and testing aggregated over each scenario. We used 5 subjects for training (2 female and 3 male), 2 for validation (1 female and 1 male) and 4 subjects for testing (2 female and 2 male). The number of video frames is the same as the number of poses (4 cameras capturing at 50Hz). The number of TOF frames can be obtained by dividing the table entries by 8 (1 sensor capturing at 25Hz).

MoCap System		DV System	
No x Sensor	10 x Vicon T40	No x Sensor	4 x Basler piA1000
Resolution	4 Megapixels	Resolution	1000x1000
Freq.	200Hz	Freq.	50Hz
Sync	hardware	Sync	hardware
TOF System		**Body Scanner**	
No x Sensor	1 x Mesa SR4000	Sensor	Vitus Smart LC3
Resolution	176x144	No. Lasers	3
Freq.	25Hz	Point Density	7dots/cm3
Sync	software	Tolerance	< 1mm

(b) Technical summary of our different sensors.



(c) Floor plan showing the capture region and the placement of the video, MoCap and TOF cameras.

Fig. 2. Overview of the data and the experimental setup. (a) Number of frames in training, validation and testing by scenario. (b) Technical specification of our sensors. (c) Schema of our capture space and camera placement.

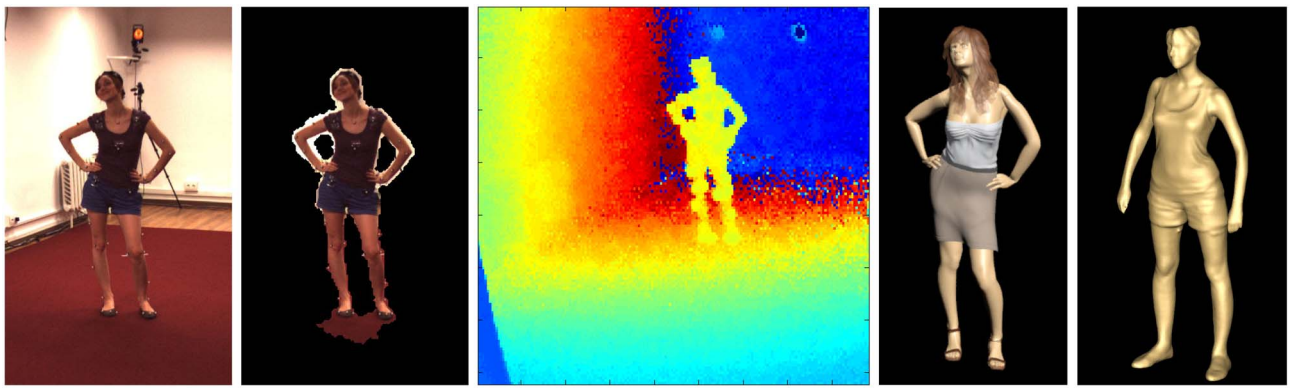


Fig. 3. Sample of the data provided in our dataset from left to right: RGB image, person silhouette (bounding box is also available), time-of-flight (depth) data (range image shown here), 3D pose data (shown using a synthetic graphics model), accurate body surface obtained using a 3D laser scanner.

of freedom to move naturally instead of being forced into a strict interpretation of the motions or poses corresponding to each task.

The dataset consists of 3.6 million different human poses collected with 4 digital cameras. Data is organized into 15 training scenarios including walking with many types of asymmetries (e.g. walking with a hand in a pocket, walking with a bag on the shoulder), sitting and lying down, various types of waiting poses and so on. The structure of the dataset is shown in Fig. 2(a), examples of images are shown in Fig. 4.

Joint Positions and Joint Angle Skeleton Representations: Common pose parametrizations considered in the literature include relative 3D joint positions (R3DJP) and kinematic representation (KR). Our dataset provides data in both parametrizations, with a full skeleton containing the same number of joints (32) in both cases. In the first case (R3DJP), the joint positions in a 3D coordinate system are provided. The data is obtained from the joint angles (provided by Vicon's skeleton fitting procedure) by applying forward kinematics on the skeleton of the subject. The parametrization is called relative because there is a specially designated joint, usually called the root (roughly corresponding to the pelvis bone position), which is taken as the center of the coordinate system, while the other joints are estimated relative to it. The kinematic representation (KR) considers the relative joint angles between limbs and is more convenient because it is invariant to both scale and body proportions. The dependencies between variables are, however, much more complex, making estimation more difficult. The process of estimating the joint angles involves non-linear optimization under joint limit constraints.

We devoted significant efforts to ensure that the data is clean and the fitting process accurate, by also monitoring the image projection errors of body joint positions. These positions were obtained based on forward kinematics, after

Fig. 4. Sample images from our dataset, showing the variability of subjects, poses and viewing angles.

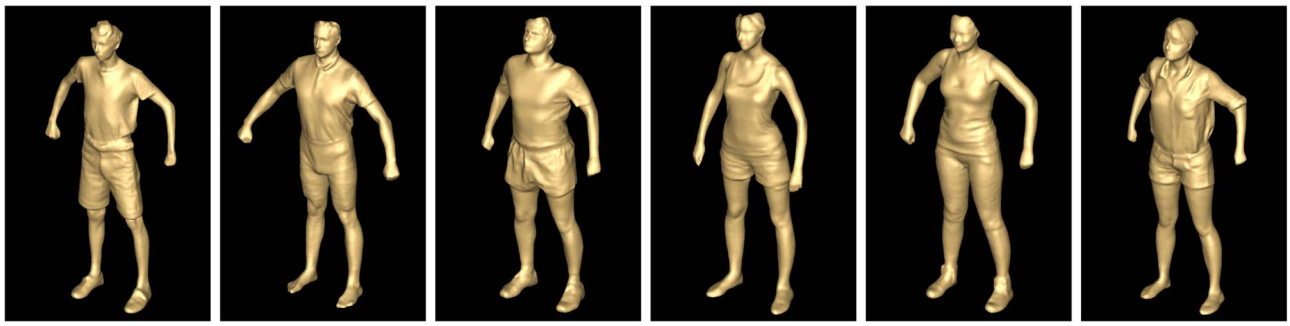


Fig. 5. High resolution meshes (body scans) of the actors involved in our experiments, illustrating the body shape variations in the dataset.

fitting, and compared against image marker tracks. Outputs were visually inspected multiple times, during different processing phases, to ensure accuracy. These representations can be directly used in independent monocular predictions or in a multi camera estimation setting. The monocular prediction dataset can be increased 4-fold by globally rotating and translating the pose coordinates to map the 4 DV cameras into a unique coordinate system (we also provide code for this data manipulation). As seen in Fig. 2(b), poses from motion capture are also available at (4-fold) faster rates compared to the images from DV cameras. Our code also provides the option to double both the image and the 3D pose data by generating their mirror symmetries. This procedure can yield 7 million images with corresponding 3D poses.

Image Processing, Silhouettes, Person Bounding Boxes and Pixel-level labels. Pixel-wise, figure-ground segmentations for all images were obtained using background models. We trained image models as mixtures of Gaussian distributions in each of the RGB and HSV color channels as well as the gradient in each RGB channel[1] (total of 3+3+2x3=12 channels). We used the background models in a graph cut framework to obtain the final figure-ground pixel labeling. The weights of the input features for the graph cut model were learned by optimizing a measure of pixel segmentation accuracy on a set of manually labeled ground truth silhouettes for a subset of images sampled from different videos. The segmentation measure we used was the standard overlap, expressed as pixel-wise intersection over the union between the hypothesis and the ground-truth silhouette. A Nelder-Mead optimization algorithm was used in order to handle non-smooth objectives.

The dataset also provides accurate bounding box annotations for people. This data was obtained by projecting the skeleton in the image and fitting a rectangular box around the projection. For accurate estimates, a separate camera calibration procedure was performed to improve the accuracy of the default one provided by the Vicon system. This extra calibration is necessary because the camera distortion parameters are not estimated by the default calibration method. The calibration data is also provided with the release of the dataset. It was obtained by positioning 30 reflective markers on the capture surface and by manually labeling them in each of the cameras with

1. Note that gradients are stable because the cameras are fixed.

Fig. 6. Sample images from our mixed reality test set. The data is challenging due to the complexity of the backgrounds, viewpoints, diverse subject poses, camera motion and occlusion.

subpixel accuracy. Models with second order radial distortion parameters were fitted to this data, separately for each of the four DV cameras. This procedure resulted in significantly improved calibration parameters, with .17 pixels mean re-projection error.

We automatically generate dense pixel-level body part labels and depth maps by using first the kinematic information to render a volumetric human model and then the camera parameters to project it in the image (see Fig. 8). For the body part labels we use a set of 24 labels, out of which 20 are associated to limbs (3 for the joints of each limb and 2 for the upper and lower bones, e.g. the arm consists of the shoulder, elbow, wrist as well as humerus and radius), 2 for the torso (chest and abdomen), 1 for the pelvis and 1 for the head.

Additional Mixed Reality Test Data: Besides creating laboratory test sets, we also focused on providing test data to cover variations in clothing and complex backgrounds, as well as camera motion and occlusion (Fig. 6). We created the mixed reality videos by inserting high quality 3D rigged animation models in real videos with realistic and complex backgrounds, good quality image data and accurate 3D pose information. The movies were created by inserting and rendering 3D models of a fully clothed synthetic character (male or female) in real videos. We are not aware of any setting of this level of difficulty in the literature. Real images may show people in complex poses, but the diverse backgrounds as well as the scene illumination and the occlusions can vary independently and represent important nuisance factors the vision systems should be robust against. Although approaches to offset such nuisance factors exist in the literature, it is difficult to evaluate their effectiveness because ground truth pose information for real images is hard to obtain. Our dataset features a component that has been especially designed to address such hard cases. This is not the only possible realistic testing scenario – other datasets in the literature [6], [7] also contain realistic testing scenarios such as different sport motions or backgrounds. Prior efforts to create mixed reality setups for training and testing 3D human pose estimation methods exist, including our own prior work [41], [42] but also [43] and more recently [44]. However, none of the prior work datasets were sufficiently large. Perhaps more importantly, the insertion of the 3D synthetic human character was not taking into account the geometry of the camera that captured the background and the one of the 3D scene (e.g. ground plane, occluders), as we do in this work.[2] The poses used for animating the models were selected directly from our laboratory test set. The Euler ZXY joint angles extracted by the motion capture system were used to create files where limb lengths were matched automatically to the models. The limb lengths were necessary in the next step, where we retargeted the captured motion data to the skeletons of the graphics models, using animation software. The actual insertion required solving for the (rigid) camera motion, as well as for its internal parameters [45], for good quality rendering. The exported camera tracks as well as the model were then imported into animation software, where the actual rendering was performed. The scene was set up and rendered using the mental ray, ray-tracing renderer, with several well-placed area lights and skylights. To improve quality, we placed a transparent plane on the ground, to receive shadows. Scenes with occlusion were also created. The dataset contains 5 different dynamic backgrounds obtained with a moving camera, a total of 7,466 frames, out of which 1,270 frames contain various degrees of occlusion. A sample of the images created is shown in Fig. 6. We see this component of **Human3.6M** as a taster. Given the large volume of motion capture data we collected, we can easily generate large volumes of mixed reality video with people having different body proportions and with different clothing, and against different real static or moving backgrounds, for both training and testing.

3 Large Scale Pose Estimation Models

We provide several large scale evaluation models with our dataset and we focus on automatic discriminative frameworks due to their conceptual simplicity and potential for

2. The insertion process involved the composition of the character silhouette sprite with the image background, with all the 3D geometric inconsistency and the image processing artifacts this can lead to.

scalability. The estimation problem is framed as learning a mapping (or an index) from image descriptors extracted over the person silhouette or its bounding box, to the pose represented based on either joint positions or joint angles. Let $\mathbf{X}_i$ be the image descriptor for frame i, $\mathbf{Y}_i$ the pose representation for frame i, and $\mathbf{f}$ (or $\mathbf{f_W}$) the mapping with parameters $\mathbf{W}$. Our goal is to estimate a model with $\mathbf{f_W}(\mathbf{X}) \simeq \mathbf{Y}$, for $\mathbf{X}$ and $\mathbf{Y}$ not seen in training. Specifically, the methods we considered are: k-nearest neighbor (kNN), linear and kernel ridge regression (LinKRR, KRR), as well as structured prediction methods based on kernel dependency estimation (KDE) [12], [46], where, for scalability reasons, we used Fourier kernel approximations [38], [47], [48]. Training such models (or any other human pose prediction method, for that matter) using millions of examples is highly non-trivial and has not been demonstrated so far in the context of such a continuous prediction problem, with structured, highly correlated outputs.

k-Nearest neighbor regression (kNN) is one of the simplest methods for learning $\mathbf{f}$ [49]. 'Training' implies storing all examples or a subset of them, in our case, due to running time constraints. Depending on the distance function used, an intermediate data structure, typically KD or cover trees [50], can be constructed during training, in order to speed-up inference at test time. These data structures, however, are dependent on the input metric and pay off mostly for problems with low input dimensionality, which is not our case. As inputs we use the χ^2 comparison metric for histograms, which is known to perform well on gradient distributions (we use pyramids of SIFT grids extracted over the person silhouette or bounding box). For vectors $\mathbf{X} = [x_1 \dots x_d]$ and $\mathbf{Y} = [y_1 \dots y_d]$, the χ^2 distance is defined as

$$\chi^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{d} \sum_l \sqrt{\frac{(x_l - y_l)^2}{x_l + y_l}} \tag{1}$$

In order to be able to run experiments within a reasonable amount of time for certain non-approximated models, we had to work with only 400K training examples, and subsample the data whenever this upper bound has been exceeded. In the experiments we used $k = 1$. In this case prediction is made by returning the stored target corresponding to the closest example from the training set under the input metric.

Kernel ridge regression (KRR) is a simple and reliable kernel method [51] that can be applied to predict each pose dimension (joint angles or joint positions) independently, with separately trained models. Parameters α_i for each model are obtained by solving a non-linear l_2 regularized least-squares problem:

$$\arg\min_{\alpha} \frac{1}{2} \sum_j \| \sum_i \alpha_i k(\mathbf{X}_j, \mathbf{X}_i) - \mathbf{Y}_j \|_2^2 + \lambda \|\alpha\|_2^2 \tag{2}$$

The problem has a closed form solution

$$\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{Y} \tag{3}$$

with $\mathbf{K}_{ij} = k(\mathbf{X}_i, \mathbf{X}_j)$ and $\mathbf{Y} = [y_1, \dots y_n]$. The weakness of the method is the cubic scaling in the training set size, because of a $n \times n$ matrix inversion. In our experiments, we choose χ^2 as our input metric and the exponential map to transform the metric into a kernel, i.e. $k(\mathbf{X}_i, \mathbf{X}_j) = \exp(-\beta \chi^2(\mathbf{X}_i, \mathbf{X}_j))$, where β is a scale parameter. This kernel is called the exponential-χ^2 kernel in the literature. Prediction is done using the rule $f_{\alpha,\beta}(\mathbf{X}) = \sum_i \alpha_i k(\mathbf{X}, \mathbf{X}_i)$.

Fourier Embeddings for Kernel Approximation. Our large scale non-linear prediction approach relies on methods to embed the data into an Euclidean space using an approximate mapping derived from the Fourier transform of the kernel. The procedure [48], relies on a theorem, due to Bochner, that guarantees the existence of such a mapping for the class of translation invariant kernels. This class contains the well-known and widely used Gaussian and Laplace kernels. The idea is to approximate a potentially infinite-dimensional or analytically unavailable kernel lifting with a finite embedding that can be computed explicitly. The approximation can be derived as an expectation in the frequency domain of a feature function ϕ which depends on the input. The expectation is computed using a density μ over frequencies, which is precisely the Fourier transform of the kernel k

$$k(\mathbf{X}_i, \mathbf{X}_j) \simeq \int_{\omega} (\phi(\mathbf{X}_i; \omega) \phi(\mathbf{X}_j; \omega)) \mu(\omega) \tag{4}$$

The existence of the measure is a key property because it allows an approximation of the integral with a Monte Carlo estimate, based on a finite sample from μ. We therefore obtain not only an explicit representation of the kernel – which is separable in the inputs, i.e., $k(\mathbf{X}_i, \mathbf{X}_j) \simeq \Phi(\mathbf{X}_i)\Phi(\mathbf{X}_j)^\top$, with $\Phi(\mathbf{X}_i) = [\phi(\mathbf{X}_i; \omega_1) \dots \phi(\mathbf{X}_i; \omega_D)]$ a vector of the $\phi(\mathbf{X}_i; \omega)$, and ω being D samples from $\mu(\omega)$ –, but at the same time we benefit from a kernel approximation guarantee, which is independent of the learning cost. The explicit Fourier feature map can then be used in conjunction with linear methods for prediction.

Linear approximations for kernel ridge regression (LinKRR) can be used to overcome the cubic computational burden of KRR while maintaining most of its non-linear predictive performance. Using the Fourier representation and standard duality arguments, one can show that equation (2) is equivalent to

$$\arg\min_{\mathbf{W}} \frac{1}{2} \sum_i \| \Phi(\mathbf{X}_i)\mathbf{W} - \mathbf{Y}_i \|_2^2 + \lambda \|\mathbf{W}\|_2^2 \tag{5}$$

This is a least squares regression model applied to non-linearly mapped data which has a closed form solution

$$\mathbf{W} = (\Phi(\mathbf{X})^\top \Phi(\mathbf{X}) + \lambda \mathbf{I}_D)^{-1} \Phi(\mathbf{X})^\top \mathbf{Y} \tag{6}$$

A matrix inversion needs to be performed in this case as well, but this time the dimension of the matrix is $D \times D$, with the input size typically much smaller than the training set ($D \ll n$). The inversion is independent of the number of examples, which makes LinKRR an attractive model for large scale training. The construction of the matrix $\Phi(\mathbf{X})^\top \Phi(\mathbf{X})$ is a linear operation in the dimension of the training set n and can be computed online with little memory consumption. Note that D is a parameter for the method and allows the trade-off between efficiency (larger D makes inversion more demanding) and performance (larger D makes the approximation more accurate). The experimental results show that often, when D is large enough, there is

little or no performance loss for many interesting kernels. To make this equivalent to the exact KRR method, we use an exponential-χ^2 kernel approximation proposed by [48].

Kernel Dependency Estimation (KDE). We also considered large-scale structured prediction models first studied in a different pose estimation context by Ionescu *et al*. [38]. The models leverage the Fourier approximation methodology for kernel dependency estimation (KDE) [46]. Standard multiple output regression models treat each dimension independently, thus ignoring correlations between targets. In many cases this simple approach works well, but for 3D human pose estimation there are strong correlations between the positions of the skeleton joints due to the physical and anatomical constraints of the human body, the environment where humans operate, or the structure and synchrony of many human actions and activities. One possibility to model such dependencies is to first decorrelate the multivariate output through orthogonal decomposition [52] based on Kernel Principal Component Analysis (KPCA) [53]. KPCA is a general framework covering both parametric kernels and data-driven kernels corresponding to non-linear manifold models or semi-supervised learning. The space recovered via kernel PCA gives an intermediate, low dimensional, decoupled representation of the outputs, and standard KRR can now be used to regress on each dimension independently. To obtain the final prediction, one needs to map from the orthogonal space obtained using Kernel PCA, and where independent KRR predictions are made, to the (correlated) pose space where the original output resides. This operation requires solving the pre-image problem [54]

$$\arg\min_{\mathbf{Y}} \|\Phi(\mathbf{X})\mathbf{W} - \Phi_{PCA}(\mathbf{Y})\|_2^2 \tag{7}$$

For certain classes of kernels, pre-images can be computed analytically, but for most kernels exact pre-image maps are not available. The general approach is to optimize (7) for the point $\mathbf{Y}$ in the target space whose KPCA projection is closest to the prediction given by the input regressor. This is a non-linear, non-convex optimization problem, but it can be solved quite reliably using gradient descent, starting from an initialization obtained from independent predictors on the original outputs. This process can be viewed as inducing correlations by starting from an independent solution.

In this case, we apply the Fourier kernel approximation methodology to both covariates and targets, in order to obtain a very efficient structured prediction method. Once the Fourier features of the targets are computed, only their dimensionality influences the complexity needed to solve for kernel PCA, and training becomes equivalent to solving a ridge regression problem to these outputs. The resulting method is very efficient, and does not require sub-sampling the data.

4 Evaluation and Error Measures

We propose several different measures to evaluate performance. Each has advantages and disadvantages, so we evaluate and provide support for all of them in order to give a more comprehensive picture of the strengths and weaknesses of different methods.[3] Let $m^{(f)}_{\mathbf{f},\mathcal{S}}(i)$ be a function that returns the coordinates of the i-th joint of skeleton $\mathcal{S}$, at frame f, from the pose estimator $\mathbf{f}$. Let also $m^{(f)}_{\mathbf{gt},\mathcal{S}}(i)$ be the i-th joint of the ground truth frame f. Let $\mathcal{S}$ be the subject specific skeleton and $\mathcal{S}_u$ be the universal skeleton. The subject specific skeleton is the one whose limb lengths correspond to the subject performing the motion. The universal skeleton has one set of limb lengths, independent of the subject who performed the motion. This allows us to obtain data in a R3DJP parametrization, which is invariant to the size of the subject.

MPJPE. Much of the literature reports *mean per joint position error*. For a frame f and a skeleton $\mathcal{S}$, MPJPE is computed as

$$E_{MPJPE}(f,\mathcal{S}) = \frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}} \|m^{(f)}_{\mathbf{f},\mathcal{S}}(i) - m^{(f)}_{\mathbf{gt},\mathcal{S}}(i))\|_2 \tag{8}$$

where $N_{\mathcal{S}}$ is the number of joints in skeleton $\mathcal{S}$. For a set of frames the error is the average over the MPJPEs of all frames.

Depending on the evaluation setup, the joint coordinates will be in 3D, and the measurements will be reported in millimeters (mm), or in 2D, where the error will be reported in pixels. For systems that estimate joint angles, we offer the option to automatically convert the angles into positions and compute MPJPE, using direct kinematics on the skeleton of the test subject (the ground truth limb lengths will not be used within the error calculation protocol).

One of the problems with this error measure is its subject specificity. Since many methods may encounter difficulties in predicting the parameters of the skeleton (e.g. limb lengths), we propose a universal MPJPE measure which considers the same limb lengths for all subjects, by means of a normalization process. We denote this error measure UMPJPE.

MPJAE. A different approach to compare poses would be to use the angles between the joints of the skeleton. We offer this possibility for methods that predict the pose in a joint angle parametrization. We call this error *mean per joint angle error* (MPJAE). The angles are computed in 3D.

$$E_{MPJAE}(f,\mathcal{S}) = \frac{1}{3N_{\mathcal{S}}}\sum_{i=1}^{3N_{\mathcal{S}}} |(m^{(f)}_{\mathbf{f},\mathcal{S}}(i) - m^{(f)}_{\mathbf{gt},\mathcal{S}}(i)) \bmod \pm 180| \tag{9}$$

In this case, the function m returns the joints angles instead of joint positions. This error is relatively unintuitive since, perceptually, not all errors should count equally. If one makes a 30° error in predicting the elbow, only one joint, the wrist, is wrongly predicted but a 30° error in the global rotation will misalign all joints.

MPJLE. The two previously proposed error measures, MPJPE and MPJAE, have two disadvantages. One issue is that they are not robust – one badly predicted joint can have unbounded impact on the error of the entire dataset. Secondly, errors that are difficult to perceive by humans can be overemphasized in the final result.

3. As the field evolves towards agreement on other metrics, not present in our dataset distribution, we plan to implement and provide evaluation support for them as well.

To address some of these observations, we propose a new error measure, the *mean per joint localization error*, that uses a perceptual tolerance parameter t

$$E_{MPJLE@t}(f,\mathcal{S}) = \frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}} \mathbb{1}_{\|m^{(f)}_{\mathbf{f},\mathcal{S}}(i) - m^{(f)}_{\mathbf{gt},\mathcal{S}}(i)\|_2 \geq t} \tag{10}$$

This error measure can be used by fixing the tolerance level using a perceptual threshold. For instance, errors below a couple of centimeters are often perceptually indistinguishable. Alternatively, errors corresponding to different tolerance levels can be plotted together. By integrating t over an interval, say $[0, 200]$, we can obtain an estimate of the average error in the same way mean average precision gives an estimate of the performance of a classifier. This error can be used also for evaluating a pose estimator that may not predict all joints, and such an estimator will be penalized only moderately. A related approach based on PCP curves has been pursued, for 2D pose estimation, in [35]. Here we differ in that we work in 3D as opposed to 2D, and we consider the joints independently as opposed to pairwise. More complex perceptual error measures beyond the ones we explore here can be envisaged. They could encode the contact between the person and its environment, including interaction with objects or contact with the ground plane. Our dataset does not contain people-object interactions but the ground plane can be easily recovered. Alternatively, in order to better understand what represents a good perceptual threshold, one can explore the degree to which people can re-enact (reproduce) a variety of human poses shown to them, in different images. This methodology is pursued in our recent work [55], and we refer the interested reader to it for details.

5 Experimental Analysis

5.1 Data Analysis

This dataset places us in the unique position of having both large amounts of data gathered from relatively unconstrained actors, with regard to stage direction, and high accuracy 3D ground-truth information. In this section we use this ground-truth information to try to gain insight into the diversity and repeatability of the poses contained in the dataset we have captured. We also take advantage of the data annotation in order to easily assess the occurrence of certain visual phenomena such as foreshortening, ambiguities and self-occlusions.

Diversity. An easy way to assess the diversity of our data is to check how many distinct poses have been obtained. We consider two poses to be distinct, if at least one joint is different than the corresponding joint from the other pose, beyond a certain tolerance t i.e. $\max_i \|m_1(i) - m_2(i)\|_2 > t$. Since our goal is to provide not only pose, but also appearance variations, poses of different subjects are considered different, independently of how similar they are in 3D. This experiment reveals that for a 100mm tolerance, 12% of the frames are distinct for a total of about 438,654 images. These figures grow to 24% or 886,409 when the tolerance is down to 50mm.

Repeatability. Pose estimation from images is a difficult problem because appearance varies not only with pose, but also with a number of "nuisance" factors like body shape and clothing. One way to deal with this problem is to isolate pose variation from all the other factors by generating a dataset of pairs of highly similar poses originating from different subjects (see Fig. 8 for examples, as well as early work on learning distance functions that preserve different levels of invariance in a hierarchical framework [42]). We compare poses using the distance between the most distant joints with a threshold at 100mm. Note that whenever a pair of similar poses is detected, temporally adjacent frames are also very similar. We eliminate these redundant pairs by clustering them in time and picking only the most similar as the representative pair. In the end, we obtain a dataset with 10,926 pairs, half of which are coming from the "Discussion", "Eating" and "Walking" scenarios.

Foreshortening. To assess the occurrence of foreshortening we consider the projections for the 3 joints of one limb (shoulder, elbow and wrist for the arms and hip, knee and ankle for the legs). We claim that such an event has happened when all these projections are very close to each other and the depth ordering of the joints is correct. In our experiments we calibrate a 20 pixel tolerance in the center of the capture surface and normalize it using the distance to the camera. We observe that although these events are rare in our dataset, they do happen. After clustering and removing redundant events, we counted 138 foreshortening events for arms and 12 for legs in the training data, and 82 and 2 respectively, for the test data. Since most of our scenarios contain mainly standing poses, foreshortening happens mostly for arms, although for the "Sitting" and "Sitting Down" scenarios they occur for legs as well (14 occurrences in total). As one might expect, the "Directions" scenario, where the subject points in different directions, has the most foreshortening occurrences (60 in total), while some scenarios like "Smoking" had none.

Ambiguities. When predicting 3D human pose from static images we are inverting an inherently lossy non-linear transformation that combines perspective projection and kinematics [18], [19]. This ambiguity makes it difficult, in the absence of priors other than the joint angle limits or the body non self-intersection constraints, to recover the original 3D pose from its projection, and the ambiguities may persist temporally [56]. The existence of monocular 3D ambiguities is well known [18], [56] but it is interesting to study to what extent these are present among the poses of a large, ecological dataset. We can assess the occurrence of ambiguities by looking at 3D and 2D ground truth pose information. We separate ambiguity events in two types. "Type 1" (T1) is an ambiguity that occurs at the level of one limb. We consider that two poses for a limb are ambiguous if the projections are closer than a threshold d_{2D} while the MPJPE is larger than some distance d_{3D}. In our experiment we use $d_{2D} = 5$ pixels and $d_{3D} = 100$ mm. These thresholds provide a large number of pairs of frames, many of which are consecutive. For a result that is easier to interpret, we group the pairs using their temporal indices and keep only one example per group. The second type of ambiguity occurs between two limbs of the same type, i.e. arms or legs. If for two different poses the projections of the joints of one limb are close to the projections of those of another limb corresponding to the second pose, while the poses

Fig. 7. Examples of visual ambiguities naturally occurring in our dataset, and automatically identified. The first two examples show foreshortening events for the left and right arm respectively. The next four images show a "Type 1" ambiguity for the right and left arm. The last two images show one "Type 2" ambiguity event: first the ambiguous leg pose and then the pose with respect to which the ambiguity was detected.

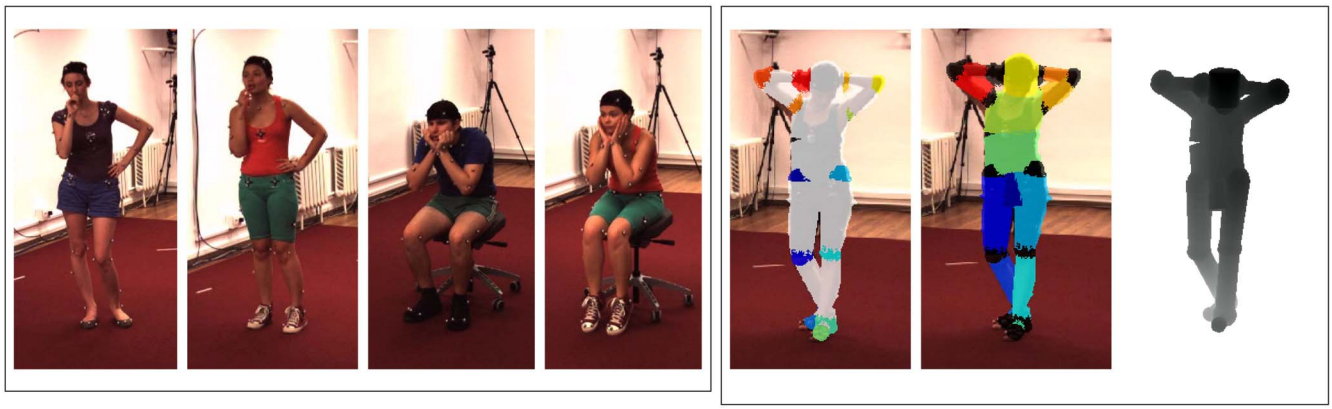


Fig. 8. Illustration of additional annotations in our dataset. The first 4 images show 2 pairs of consistent poses found from different subjects. The last 3 images are obtained using 3D pose data. We render a simple geometric model and project it in the image to obtain additional annotation: joint visibility, body part support in the image and even rough pixel-wise 3D information. This allows us to detect joint occlusion, partial limb occlusion and obtain depth discontinuities.

are still relatively consistent, i.e. MPJPE is not too large, then we consider to have a "Type 2" (T2) ambiguity. The constraint on MPJPE is added to remove forward-backward flips which are the most likely cause of similar projections for different limbs. Examples of both types of ambiguities are given in Fig. 7. Notice however that the presence of a subset of ambiguities in our captured data does not imply that such ambiguities and their number would immediately correlate to the ones obtained by an automatic monocular pose estimation system–we know that a larger set of geometric ambiguities exists, *ex-ante*. The question is to what extent the pose estimation system can be made 'not see them' using image constraints, prior knowledge, or information from the environment and the task. The results are summarized in Table 1.

Self-occlusion. Unlike 2D pose estimation datasets, our data does not directly provide information about joint and limb visibility. This can be computed using the available data by considering the 3D pose information and using it to render a simple geometric model which can then be projected onto the image. In Fig. 8, we show examples of the *joint locations* from which body part label visibility can be easily obtained. Moreover, we can obtain *dense part labels* from which part visibility can be derived and *depth information* which can be used to label depth discontinuity edges. All these detailed annotations are very difficult to obtain in general, and are highly relevant in the context of human pose estimation.

5.2 Prediction Experiments

In this section we provide quantitative results for several methods including nearest neighbors, regression and large-scale structured predictors. Additionally, we evaluate subject and activity specific models, as well as general

TABLE 1
Summary of the Results for Our Type 1 (T1) and Type 2 (T2) Ambiguity Experiments Showing Counts of Distinct Ambiguity Events by Scenario, in Our Dataset

	T1LArm	T1RArm	T1LLeg	T1RLeg	T2Legs
Directions	1235	2512	3022	3219	425
Discussion	7501	9503	8226	6167	881
Eating	2451	3130	3277	3100	175
Greeting	1507	2099	2392	2066	228
Phone Talk	2255	3154	3316	3045	191
Posing	1767	2468	2431	2145	117
Buying	922	1311	1205	962	96
Sitting	2398	3220	3508	3693	4
Sitting Down	2200	2996	3270	3407	66
Smoking	2109	3574	3660	3320	232
Taking Photo	1096	1407	1831	1611	109
Waiting	2893	3820	4387	3353	265
Walking	2407	3017	3266	2225	965
Walking Dog	1142	1395	1592	1468	298
Walking Pair	925	1406	1828	1778	366

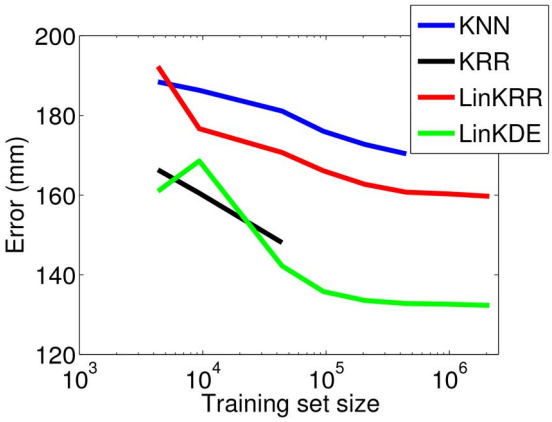


Fig. 9. Evolution of the test error as a function of the training set size. Larger training sets offer important performance benefits for all models we have tried.

models trained on the entire dataset. We also study the degree of success of the methodology in more challenging situations, like the ones available in our mixed reality dataset.

Image Descriptors. For silhouette and person bounding box description, we use a pyramid of grid SIFT descriptors with 3 levels (2x2, 4x4 and 8x8) and 9 orientation bins. Variants of these features have been shown to work well on previous datasets (e.g. HumanEva [4])[14], [41] and we show that these are quite effective even in this more complex setting. Since both background subtraction (BS) and bounding box (BB) localization of our subjects are provided, we performed the experiments using both features extracted over the entire bounding box, and using descriptors where the BS mask is used, in order to filter out some of the background.

Pose Data. Our 3D pose data is mapped to the coordinate system of one of the 4 cameras, and all predictions are performed in that coordinate system. When we report joint position errors, the root joint of the skeleton is always in the center of the coordinate system used for prediction. Errors are reported mostly in mm using MPJPE. Sometimes we use MPJAE which reports errors in angle degrees. Human poses are represented using a skeleton with 17 joints. This limitation of the number of joints helps discard the smallest links associated to details for the hands and feet, going as far down the kinematic chain to only reach the wrist and the ankle joints.

Training set size. We first studied the manner in which test errors vary with the size of the training set (Fig. 9). Due to the limited memory and computational resources available, the largest exact KRR model was learnt using 50,000 samples, and the largest kNN model was based on 400,000 samples. The results show that our best performing approximate non-linear model, LinKDE, capitalizes on a 2 orders of magnitude increase in training set size by reducing the test error by roughly 20%. This experiment clearly shows the potential impact of using **Human3.6M** in increasing the accuracy of pose estimation models. In fact, the potential for research progress is significantly vaster, as more sophisticated models with increased capacity, beyond our baselines here, can be used, with two orders of magnitude more data than in the largest available dataset.

Results. Linear Fourier methods use input kernel embeddings based on 15,000-dimensional random feature maps (corresponding to exponentiated χ^2), and 4000-d output kernel embedding (corresponding to Gaussian kernels). Typical running times on an 8 core PC with the full dataset include 16 hours for testing kNN models (with a training set subsampled to 300K examples), 1h for training and 12h for testing KRR (40K example training set). For the full training set of 2.1 million examples where only linear approximations to non-linear models can be effectively applied, training LinKRR takes 5h and testing takes about 2h. LinKDE takes about 5h to train and 40h to test. Code for all of the methods is provided on our website as well.

Several training and testing model scenarios were prepared. The simplest one considers data from each subject separately (we call this Subject Specific Model or SSM). The motions for our 15 scenarios are each captured in 2 trials which are used for training and validation, respectively. A set of 2 motions from each subject were reserved for testing (these were data captured in distinct motions performed by the subjects, not subsampled from single training sequences). The setup includes different poses that appear in the 15 training scenarios (one involves sitting, the second one does not). This type of experiment was designed to isolate the pose variability from the body shape and clothing variability. A second, more challenging scenario, considers prediction with a model trained on a set of 7 fixed training subjects (5 for training, and 2 for validation) and tested on the remaining 4 subjects on a per motion basis (we call it Activity Specific Model or the ASM). Finally, we used a setup where all motions are considered together using the same split among subjects (our General Model, GM).

We first tested the baseline methods on the simplest setup, SSM (see Table 2). We noticed a difference between results obtained using background subtraction (BS) and bounding box (BB) inputs with a slight edge to BB. This is not entirely surprising when considering, for instance, examples involving sitting poses. There, the presence of the chair makes background subtraction very difficult and that affects the positioning of the object within the descriptor's coordinate system. This problem only affects our background subtraction data since the bounding boxes are computed from the joint projections alone.

In our second setup we tested our models on each motion separately. These are referred to as Activity Specific Models (ASM). We noticed that errors are considerably higher both because of the large size of our test set and the significant subject body variation introduced. Our 'sitting down' motion is one of the most challenging. It consists of subjects sitting on the floor in different poses. This scenario is complex to analyze because of the high rate of self-occlusion, as well as the bounding box aspect ratio changes. It also stretches the use of image descriptors extracted on regular grids, confirming that, while these may be reasonable for standing poses or pedestrians, they are not adequate for general human motion sensing. The other 'sitting' scenario in the dataset is challenging too due to the use of external objects, in this case a chair. The 'taking photo' and 'walking dog' motions are also difficult because of bounding box variations, and because they are less repeatable and more liberty was granted to the actors performing them. Overall,

TABLE 2
Results of the Different Methods, Corresponding to the Subject Specific Modeling (SSM) Setup, and for All Training Subjects in the Dataset. kNN Indicates Nearest Neighbor (k=1), KRR Is Kernel Ridge Regression, and LinKRR Represents a Linear Fourier Approximation of KRR. LinKDE Is the Linear Fourier Approximation Corresponding to a Structured Predictor Based on Kernel Dependency Estimation (KDE). Errors Are Given in mm, Using the MPJPE Metric

method	mask	S1	S7	S8	S9	S11	S5	S6
kNN	BB	118.93	129.60	74.90	113.31	127.98	132.00	155.65
kNN	BS	127.91	112.19	63.27	108.68	132.96	113.65	139.35
KRR	BB	99.96	96.41	58.94	95.75	106.50	108.35	117.90
KRR	BS	107.96	100.66	58.19	97.73	114.84	112.18	114.60
LinKRR	BB	114.98	114.30	81.69	119.55	126.35	128.00	140.86
LinKRR	BS	125.46	122.89	82.09	122.29	136.81	134.84	141.01
LinKDE	BB	94.07	93.63	55.32	91.80	97.25	96.35	113.80
LinKDE	BS	96.13	93.51	51.95	89.54	100.96	105.89	102.74

TABLE 3
Comparison of Predictors for the Activity Specific Setting (ASM), on the Test Set (Including S10). kNN Indicates Nearest Neighbor (k=1), KRR kernel Ridge Regression, LinKRR Is a Linear Fourier Approximation of KRR, and LinKDE Is the Linear Fourier Model for a Structured Predictor Based on Kernel Dependency Estimation (KDE). Errors Are Given in mm, Using the MPJPE Metric

method	mask	Directions	Discussion	Eating	Greeting	Phone Talk	Posing	Buying	Sitting
kNN	BB	154.23	151.18	136.23	165.94	147.51	175.58	180.82	194.09
kNN	BS	166.28	163.68	132.19	188.33	145.89	199.89	174.38	174.67
KRR	BB	118.96	116.77	109.65	128.51	123.05	136.23	153.55	176.90
KRR	BS	130.04	124.96	118.60	140.73	125.35	152.21	157.45	159.30
LinKRR	BB	123.67	121.23	116.09	136.77	130.60	142.40	165.14	180.69
LinKRR	BS	136.07	132.33	123.90	149.99	132.83	158.98	162.36	168.12
LinKDE	BB	115.79	113.27	99.52	128.80	113.44	131.01	144.89	160.92
LinKDE	BS	124.19	117.44	93.01	138.90	111.40	145.43	136.94	139.29

method	mask	Sitting Down	Smoking	Taking Photo	Waiting	Walking	Walking Dog	Walking Pair
kNN	BB	209.06	161.22	234.05	176.16	167.00	239.38	180.91
kNN	BS	237.05	169.41	247.57	193.78	158.27	216.53	189.80
KRR	BB	184.58	120.19	182.50	139.66	129.13	183.27	143.19
KRR	BS	213.72	130.47	197.21	150.39	119.28	175.34	150.43
LinKRR	BB	204.62	128.62	194.32	144.54	133.49	191.92	147.87
LinKRR	BS	231.57	139.88	208.14	157.92	126.90	185.64	156.33
LinKDE	BB	172.98	114.00	183.09	138.95	131.15	180.56	146.14
LinKDE	BS	203.10	118.37	197.13	146.30	115.28	166.10	153.59

TABLE 4
Results of Our GM Setup, with Models Estimated Based on Data from All Subjects and Activities in the Training Set, and Evaluated on the Full Test Set, Including S10. LinKRR (LKRR) and LinKDE (LKDE) Are Kernel Models Based on Random Fourier Approximations Trained and Tested on 2.1M and 1.4M Poses Respectively. The Exact KRR Results are Obtained by Using a Subset of Only 40,000 Human Poses Sampled from the Training Set. The Results for Joint Positions Are in mm Using MPJPE and the Results with Angles Are in Degrees Computed Using MPJAE

Joint Positions							
BB				BS			
kNN	KRR	LKRR	LKDE	kNN	KRR	LKRR	LKDE
172.12	138.85	150.73	127.92	182.79	151.73	162.51	137.98

Joint Angles							
BB				BS			
kNN	KRR	LKRR	LKDE	kNN	KRR	LKRR	LKDE
18.28	13.83	13.86	13.68	17.75	13.83	13.92	13.74

we feel that the dataset offers a good balance between somewhat 'easier' settings, as well as moderately difficult and challenging ones, making it a plausible benchmark for testing new and improved features, models or algorithms.

Due to certain privacy concerns, we have decided to withhold the images of one of our testing subjects, S10. However, we make all the other data associated to this subject available, including silhouettes, bounding boxes as well as corresponding image descriptors. In this article we report results for both ASM and GM models, including S10 in the test set (Tables 3 and 4). In the future, as other features are developed by external researchers or by us, we will strive to compute those and make them available for download for S10, too. Our evaluation server allows error evaluation on the test set, both including and excluding S10.

The MPJLE measure gives insight into the success and failures of our tested methods (Fig. 10). For the 'Eating' ASM, one of the easier activities, LinKDE correctly predicts, on average, 14 out of 17 joints, at 150mm tolerance. In contrast KRR and LinKRR correctly predict only 12 joints at the same level of tolerance.

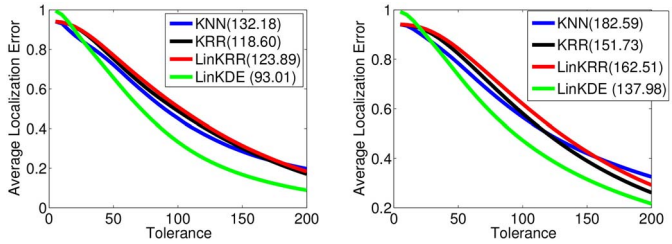


Fig. 10. Proposed perceptual MPJLE error measure. In the left plot we show the MPJLE for the 'Eating' ASM. We compare all our predictors, where features are computed on input segments given by background subtraction. The right plot shows similar results for the background subtraction GM. MPJPE errors for each model are given in parentheses. The results are computed using the full test set (including subject S10).

TABLE 5
Pose Estimation Error for Our Mixed Reality Dataset, Obtained with Moving Cameras, and Under Challenging Non-Uniform Backgrounds and Occlusion (See Fig. 6). The Errors Are Computed Using MPJAE and Do Not Include the Global Rotation. LinKRR (Here LKRR) and LinKDE (LKDE) Are Linear Fourier Approximation Methods. The Models Were Trained on the Data Captured in the Laboratory and We Tested on the Mixed-Reality Sequences. For ASM, We Used the Model Trained on Motions of the Same Type as the Test Motion. The Results Are Promising but also Show Clear Scope for Feature Design and Model Improvements (the Methods Shown Do Not Model or Predict Occlusion Explicitly)

	Activity Specific Model (ASM)								General Model (GM)							
	BB				BS				BB				BS			
	kNN	KRR	LKRR	LKDE	kNN	KRR	LKRR	LKDE	kNN	KRR	LKRR	LKDE	kNN	KRR	LKRR	LKDE
MR1	25.83	20.09	19.81	19.85	21.23	18.04	18.58	18.47	24.99	20.64	20.11	19.61	22.76	19.44	18.82	18.45
MR2	20.40	16.40	17.27	16.81	19.43	16.16	16.28	15.28	20.29	18.00	18.02	16.53	19.21	16.64	16.91	15.51
MR3	24.08	20.75	21.53	21.20	25.60	22.40	21.89	21.84	23.40	20.91	21.95	21.23	25.78	21.75	22.02	21.75
MR4	25.69	19.86	20.64	20.26	22.40	19.67	20.12	19.35	26.31	21.53	20.76	19.89	23.36	20.08	19.90	19.03
MR5	19.36	17.13	17.54	17.31	19.04	16.52	16.72	16.51	21.50	20.87	21.49	19.99	25.57	20.85	22.14	19.33
MR6	20.47	18.49	19.55	18.81	20.95	18.07	18.11	17.36	26.26	22.58	22.29	20.58	23.00	20.16	20.24	18.79
MR7	19.03	16.83	17.13	14.70	18.35	13.87	15.52	13.88	20.28	19.12	17.78	16.21	20.21	19.03	18.71	16.53

Our final evaluation setup is the one where models are trained based on all motions from all subjects. Due to the size of the dataset, this is a highly non-trivial process and very few existing methodologies can handle it. We refer to this as the general motion (GM) setup and show the results in Table 4. The models we have tested appear not to be able yet to effectively leverage the structure in the data, but it is encouraging that linear Fourier approximations to non-linear methods can be applied on such large datasets with promising results, and within a reasonable time budget. Future research towards better image descriptors, improved modeling of correlations among the limbs of the human body, or the design of large scale learning methods should offer new insights into the structure of the data and should ultimately improve the 3D prediction accuracy.

Mixed Reality Results. We use models trained on our laboratory data and test on mixed reality data. The results are given in Table 5. The test videos are named mixed-reality (MR) 1 to 7. We consider 2 scenarios: one using the ASM of the activity from which the test video was generated and one using the GM. In this experiment we use MPJAE (in degrees) and, for technical reasons, ignore the error corresponding to the global rotation. The ASM results are in general better than the GM results reflecting a more constrained prediction problem. As expected, BS results are better than BB results, showing that benefits from slightly more stable training features, observed for BB in the laboratory setting, are offset by the contribution of real background features.

6 Conclusion

We have introduced a large scale dataset, **Human3.6M** containing 3.6 million different 3D articulated poses captured from a set of professional men and women actors. **Human3.6M** complements the existing datasets with a variety of human poses typical of people seen in real-world environments, and provides synchronized 2D and 3D data (including time of flight, high quality image and motion capture data), accurate 3D human models (body surface scans) of the actors, and mixed reality settings for performance evaluation under realistic backgrounds, correct 3D scene geometry, and occlusion. We also provide studies and evaluation benchmarks based on discriminative pose prediction methods. Our analysis includes not only nearest neighbor or standard linear and non-linear regression methods, but also advanced structured predictors and large-scale approximations to non-linear models based on explicit Fourier feature maps. The ability to train complex approximations to non-linear models on millions of examples opens up possibilities to develop alternative feature descriptors and correlation kernels, and to test them seamlessly, at large scale. We show that our full dataset delivers important performance benefits compared to smaller equivalent datasets, but also that significant space for improvement exists. The data, as well as the large-scale structure models, the image descriptors, as well as the visualization and software evaluation tools we have developed are freely available online, for academic use. We hope that **Human3.6M** and its associated tools will stimulate further research in computer vision, machine learning, and will help in the development of improved 3D human sensing systems that can operate robustly in the real world.

Acknowledgment

This work was supported in part by CNCS-UEFICSDI, under PNII RU-RC-2/2009, PCE-2011-3-0438, and CT-ERC-2012-1. We thank our colleague S. Cheran, for support with the Web server. We are also grateful to F. Li for helpful discussions and for experimental feedback on Fourier models.

References

[1] T. Moeslund, A. Hilton, V. Kruger, and L. Sigal, *Visual Analysis of Humans: Looking at People*. London, U.K.: Springer Verlag, 2011.
[2] B. Rosenhahn, R. Klette, and D. Metaxas, *Human Motion, Understanding, Modelling, Capture and Animation,* vol. 36. Dordrecht, Netherlands: Springer Verlag, 2008.
[3] C. Sminchisescu, "3D human motion analysis in monocular video: Techniques and challenges," in *Human Motion.* Dordrecht, Netherlands: Springer, 2008.
[4] L. Sigal, A. Balan, and M. Black, "HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vis.*, vol. 87, no. 1, pp. 4–27, Mar. 2010.
[5] M. Yamada, L. Sigal, and M. Raptis, "No bias left behind: Covariate shift adaptation for discriminative 3D pose estimation," in *Proc. Eur. Conf. Computer Vision*, Florence, Italy, 2012.
[6] D. Vlasic *et al.*, "Practical motion capture in everyday surroundings," in *Proc. SIGGRAPH*, New York, NY, USA, 2007.
[7] G. Pons-Moll *et al.*, "Outdoor human motion capture using inverse kinematics and von mises-fisher sampling," in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain, 2011.

[8] *CMU HMC* [Online]. Available: http://mocap.cs.cmu.edu/search.html

[9] C. Sminchisescu, L. Bo, C. Ionescu, and A. Kanaujia, *Feature-based Human Pose Estimation, in Guide to Visual Analysis of Humans: Looking at People*. London, UK: Springer, 2011.

[10] R. Rosales and S. Sclaroff, "Learning body pose via specialized maps," in *Proc. NIPS*, Cambridge, MA, USA, 2002.

[11] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 44–58, Jan. 2006.

[12] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "BM^3E: Discriminative density propagation for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 2030–2044, Nov. 2007.

[13] L. Bo and C. Sminchisescu, "Structured output-associative regression," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Miami, FL, USA, 2009.

[14] L. Bo and C. Sminchisescu, "Twin Gaussian processes for structured prediction," *Int. J. Comput. Vision*, vol. 87, no. 1–2, pp. 28–52, Mar. 2010.

[15] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Hilton Head Island, SC, USA, 2000.

[16] H. Sidenbladh, M. Black, and D. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. European Conf. Computer Vision*, Dublin, Ireland, 2000.

[17] C. Sminchisescu and B. Triggs, "Covariance-scaled sampling for monocular 3D body tracking," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2001.

[18] C. Sminchisescu and B. Triggs, "Kinematic jump processes for monocular 3D human tracking," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Madison, WI, USA, 2003, pp. 69–76.

[19] C. Sminchisescu and B. Triggs, "Building roadmaps of minima and transitions in visual models," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 81–101, 2005.

[20] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard, "Tracking loose-limbed people," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2004.

[21] R. Li, M. Yang, S. Sclaroff, and T. Tian, "Monocular tracking of 3D human motion with a coordianted mixture of factor analyzers," in *Proc. Eur. Conf. Computer Vision*, Graz, Austria, 2006.

[22] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, San Francisco, CA, USA, 2010.

[23] J. Gall, B. Rosenhahn, T. Brox, and H. Seidel, "Optimization and filtering for human motion capture: A multi-layer framework," *Int. J. Comput. Vision*, vol. 87, pp. 75–92, Mar. 2010.

[24] L. Sigal, A. Balan, and M. J. Black, "Combined discriminative and generative articulated pose and non-rigid shape estimation," in *Proc. NIPS*, 2007, pp. 1337–1344.

[25] Y. Liu *et al.*, "Markerless motion capture of multiple characters using multi-view image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2720–2735, Nov. 2013.

[26] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black, "Drape: Dressing any person," *ACM Trans. Graphics*, vol. 31, no. 4, pp. 35:1–35:10, Jul. 2012.

[27] D. Anguelov *et al.*, "SCAPE: Shape completion and animation of people," in *Proc. SIGGRAPH*, New York, NY, USA, 2005.

[28] *INRIA 4D* [Online]. Available: http://4drepository.inrialpes.fr/pages/home

[29] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast pose estimation with parameter sensitive hashing," in *Proc. IEEE Int. Conf. Computer Vision*, Nice, France, 2003.

[30] A. Agarwal and B. Triggs, "3D human pose from silhouettes by relevance vector regression," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2004.

[31] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3d human pose annotations," in *Proc. IEEE Int. Conf. Computer Vision*, Kyoto, Japan, 2009.

[32] D. Ramanan and C. Sminchisescu, "Training deformable models for localization," in *Proc. CVPR*, Washington, DC, USA, 2006.

[33] V. Ferrari, M. Marin, and A. Zisserman, "Pose seach: Retrieving people using their pose," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Miami, FL, USA, 2009.

[34] B. Sapp, A. Toshev, and B. Taskar, "Cascaded models for articulated pose estimation," in *Proc. European Conf. Computer Vision*, Heraklion, Greece, 2010, pp. 406–420.

[35] M. Eichner and V. Ferrari, "We are family: Joint pose estimation of multiple persons," in *Proc. European Conf. Computer Vision*, Heraklion, Greece, 2010, pp. 228–242.

[36] Y. Wang, D. Tran, and Z. Liao, "Learning hierarchical poselets for human parsing," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Washington, DC, USA, 2011, pp. 1705–1712.

[37] T. Shiratori, H. S. Park, L. Sigal, Y. Sheikh, and J. Hodgins, "Motion capture from body-mounted cameras," in *Proc. SIGGRAPH*, New York, NY, USA, 2011.

[38] C. Ionescu, F. Li, and C. Sminchisescu, "Latent structured models for human pose estimation," in *Proc. IEEE Int. Conf. Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 2220–2227.

[39] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real time motion capture using a single time-of-flight camera," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, San Francisco, CA, USA, 2010, pp. 755–762.

[40] J. Shotton *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2011, pp. 119–135.

[41] C. Sminchisescu, A. Kanaujia, and D. Metaxas, "Learning joint top-down and bottom-up processes for 3D visual inference," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Washington, DC, USA, 2006, pp. 1743–1752.

[42] A. Kanaujia, C. Sminchisescu, and D. Metaxas, "Semi-supervised hierarchical models for 3D human pose reconstruction," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Minneapolis, MN, USA, 2006, pp. 1–8.

[43] A. Agarwal and B. Triggs, "A local basis representation for estimating human pose from cluttered images," in *Proc. ACCV*, Hyderabad, India, 2006, pp. 50–59.

[44] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Providence, RI, USA, 2012, pp. 3178–3185.

[45] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, R. Hartley and A. Zisserman, Eds. Cambridge, U.K.: Cambridge University Press, 2000.

[46] C. Cortes, M. Mohri, and J. Weston, "A general regression technique for learning transductions," in *Proc. Int. Conf. Machine Learning*, New York, NY, USA, 2005, pp. 153–160.

[47] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. NIPS*, 2007, pp. 1177–1184.

[48] F. Li, G. Lebanon, and C. Sminchisescu, "Chebyshev approximations to the histogram χ^2 kernel," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Providence, RI, USA, 2012, pp. 2424–2431.

[49] G. Shakhnarovich, T. Darrell, and P. Indyk, *Nearest-Neighbors Methods in Learning and Vision: Theory and Practice*. Cambridge, MA, USA: MIT Press, 2006.

[50] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *Proc. Int. Conf. Machine Learning*, Pittsburgh, PA, USA, 2006, pp. 97–104.

[51] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Ann. Statist.*, vol. 36, no. 3, pp. 1171–1220, Jan. 2008.

[52] J. Weston, O. Chapelle, A. Elisseeff, B. Schölkopf, and V. Vapnik, "Kernel dependency estimation," in *Proc. NIPS*, 2002, pp. 873–880.

[53] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, Jul. 1998.

[54] G. Bakir, J. Weston, and B. Scholkopf, "Learning to find pre-images," in *Proc. NIPS*, 2004.

[55] E. Marinoiu, D. Papava, and C. Sminchisescu, "Pictorial human spaces: How well do humans perceive a 3D articulated pose?" in *Proc. IEEE Int. Conf. Computer Vision*, 2013, pp. 1289–1296.

[56] C. Sminchisescu and A. Jepson, "Variational mixture smoothing for non-linear dynamical systems," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Washington, DC, USA, 2004, pp. 608–615.

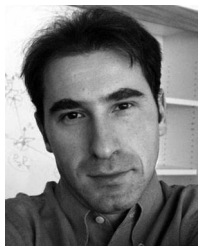
[57] M. Brubaker and D. Fleet, "The kneed walker for human pose tracking," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1–8.

[58] R. Urtasun, D. Fleet, A. Hertzmann, and P. Fua, "Priors for people tracking in small training sets," in *Proc. IEEE Int. Conf. Computer Vision*, Washington, DC, USA, 2005, pp. 403–410.
[59] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan, "Computational studies of human motion: Part 1, tracking and motion synthesis," *Found. Trends Comput. Graph. Vis.*, vol. 1, no. 2–3, pp. 77–254, 2006.
[60] R. Plankers and P. Fua, "Articulated soft objects for video-based body modeling," in *Proc. IEEE Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001, pp. 394–401.
[61] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Comput. Graph. Appl.*, vol. 27, no. 3, pp. 21–31, May 2007.

Catalin Ionescu is a PhD candidate at the University of Bonn, Bonn, Germany. As part of DFH-UFA, a Franco-German joint degree program, Catalin has received the Diplome d'ingenieur from the Institut National de Sciences Appliques de Lyon and the Diplom für Informatik from the University of Karlsruhe, Karlsruhe, Germany. His current research interests include large scale machine learning and computer vision with emphasis on 3D pose estimation.

Dragos Papava received the bachelor's degree in computer science from Politehnica University of Bucharest, Bucharest, Romania, and the master's degree in computer graphics, multimedia, and virtual reality at the same university. His current research interests include computer graphics, GPU computing, and image processing.

Vlad Olaru received the BS degree from the "Politehnica" University of Bucharest, Bucharest, Romania, the MS degree from Rutgers, the State University of New Jersey, New Brunswick, NJ, USA, and the PhD degree from the Technical University of Karlsruhe, Karlsruhe, Germany. His current research interests include distributed and parallel computing, operating systems, real-time embedded systems, and high-performance computing for large-scale computer vision programs. His doctorate concentrated on developing kernel-level, single system image services for clusters of computers. He was a key person in several EU-funded as well as national projects targeting the development of real-time OS software to control the next generation of 3D intelligent sensors, real-time Java for multi-core architectures, servers based on clusters of multi-core architectures.



Cristian Sminchisescu has received the doctorate in computer science and applied mathematics with an emphasis on imaging, vision, and robotics at INRIA, France, under an Eiffel Excellence Doctoral Fellowship, and has done post-doctoral research at the Artificial Intelligence Laboratory, University of Toronto, Toronto, ON, Canada. He is a member with the program committees of the main conferences in computer vision and machine learning (CVPR, ICCV, ECCV, NIPS, AISTATS), Area Chair for ICCV07-13, and an associate editor of IEEE PAMI. He has given over 100 invited talks and presentations and has offered tutorials on 3D tracking, recognition and optimization at ICCV and CVPR, the Chicago Machine Learning Summer School, the AERFAI Vision School in Barcelona and the Computer Vision Summer School (VSS) in Zurich. His current research interests include area of computer vision (3D human pose estimation, semantic segmentation) and machine learning (optimization and sampling algorithms, structured prediction, and kernel methods).